Breast Cancer Data Analysis and Predictive Modeling

By Krishnam Vajra

Exploratory Data Analysis

- The Dataset has Medical Records of **198** breast cancer patients.
- There are **35** medical features that help us in analysing whether a patient has a tumor that is recurrent or not
- This is a Binary Classification problem with classes: Recurrent Tumor(1) and Non-recurrent Tumor(0).
- The Dataset is quite imbalanced as the number of patients with Recurrent tumor are **47** and number of patients with non recurrent tumor are **151**.
- There are **4** missing values marked as ‘?’ in ‘Lymph_Node_Status’ feature.
- There is a feature which provides the value of the recurrence time for patients.

Observations

- As we see, The dataset is imbalanced. Recurrent classes are 47 and Nonrecurrent classes are 151.
- The number of positive axillary lymph nodes of different for different patients. with largest number of lymph nodes to be as high as 27 where as few patients have no lymph nodes at all.
- We can observe that the maximum time of recurrence of tumor is 125 and lowest is 1 units of time.
- The correlation matrix gives the pairwise correlation between all the columns.
- We can clearly observe that the time of recurrence is poorly correlated with almost all the features.
- It is important to notice that there is strong correlation between Smoothness, compactness, concavity of cells with fractal dimension of cells.
- There is strong correlation between features that are related to radius, perimeter, texture area of cell which is quite intuitive.

The whole task is performed in following steps

- Data Loading.
- Data Cleaning.
- Exploratory Data Analysis.
- Predictive Modeling.
- Tackling the Imbalanced data.
- Modeling with Deep Learning.

Classification Model results.

- **Logistic Regression**: accuracy: 0.85, f1 score: 0.67, confusion matrix: [[42 0], [9 9]]
- **KNN**: accuracy: 0.7, f1 score: 0.31, confusion matrix: [[38 4], [14 4]]
- **Decision Trees**: accuracy: 0.65, f1 score: 0.46, confusion matrix: [[30 12], [9 9]]
- **Random Forest**: accuracy: 0.73, f1 score: 0.20, confusion matrix: [[42 0], [16 2]]
- **Naive Bayes**: accuracy: 0.7, f1 score: 0.50, confusion matrix: [[33 9], [9 9]]
- **SVM**: accuracy: 0.71, f1 score: 0.11, confusion matrix: [[42 0], [17 1]]

As it is an imbalanced dataset, using f1-score and confusion matrix over accuracy will be more prudent.

As we can see that Logistic regression has performed well. But the f1-score, number of false negatives are still not satisfactory. As the dataset is imbalanced the false negatives are high.

So, We try to tackle the issue by doing over sampling of minority classes. After This technique the results are:

Logistic Regression with over sampling: array([[36, 10], [6, 39]]) We have not yet achieved optimal false negatives value. So, We perform an exhaustive grid search to find the best hyper parameters.

The best model we found is the one with following parameters:

```
{'C': 10.0, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'max_iter':
100, 'multi_class': 'ovr', 'n_jobs': 1, 'penalty': 'l1', 'random_state': None,'solver': 'liblinear', 'tol':
0.0001, 'verbose': 0, 'warm_start': False}
```

Its results:

Confusion matrix: [[34, 12], [3, 42]]), ROC: 0.866, f1_score: 0.84848, accuracy: 0.835

This model has performed quite well when it is compared to earlier models.

Now, We move on to Deep Learning Architecture.

Deep Learning Architecture Results

Confusion matrix:([[35, 11], [0, 45]]), F1-score: 0.89, Accuracy: 87.91

We can see the false negatives have come down to 0 which is ideal and F1-score and accuracies significantly better.

TASK 2:

We have used SVR(Support Vector regression model to find values of recurrence time for patients with ‘R’ class.

The RMSE Value of the model is 25.64.

Even better results can be achieved if we have more data as we currently have only 198 medical records of patients.

Thank you.